CRF-based Statistical Learning of Japanese Accent Sandhi for Developing Japanese Text-to-Speech Synthesis Systems

Nobuaki Minematsu†, *Ryo Kuroiwa*‡, *Keikichi Hirose*‡, *Michiko Watanabe*†

† Graduate School of Frontier Sciences, The University of Tokyo
‡ Graduate School of Information Science and Technology, The University of Tokyo
{mine,kuroiwa,hirose,watanabe}@gavo.t.u-tokyo.ac.jp

Abstract

In Japanese, every content word has its own H/L pitch pattern when it is uttered isolatedly, called accent type. In a TTS system, this lexical information is usually stored in a dictionary and it is referred to for prosody generation. When converting a written sentence to speech, however, this lexical H/L pattern is often changed according to the context, known as word accent sandhi. This accent change is troublesome for speech synthesis researchers because it is difficult even for native speakers to describe explicitly what kind of mechanism is working for the change although young Japanese learn the mechanism without trouble. For developing a good Japanese TTS system, this implicit and phonological knowledge has to be built in the system. In our previous study [1], we developed a rule-based module for the accent sandhi but it is true that it produced an unignorable number of errors. In this paper, the development of a corpus-based module is described using Conditional Random Fields (CRFs) to predict the change. Although the new module shows the better performance for the prediction than the previous rule-based module, the new module is tuned further by integrating the rule-based knowledge acquired in the previous study.

1. Introduction

Several functions, such as text analysis, grapheme to phoneme conversion, and speech waveform generation, need to be developed to realize a TTS conversion system. Among them, the generation of prosodic features from an input text is very important and requires a sophisticated process, since no information on prosody is directly given in the text. Especially in the case of Japanese, the control of fundamental frequency (henceforth F_0) movement is crucial to achieve the high quality in the synthetic speech. In order to realize a good prosody control, the location of the accent nucleus should be adequately estimated for each accentual phrase as well as the boundaries of prosodic clauses (breath groups), prosodic phrases, and accentual phrases.

An accentual phrase of Japanese is often composed of two words or more, typically a content word followed by a function word. Although all the content words (and some function words) have their own accent nucleus position as their lexical attribute, the accent nucleus of an accentual phrase often shifts due to the accent sandhi. This accent shift has to be correctly predicted in TTS conversion. Some rules of the accent sandhi can be found in some accent dictionaries such as [2] but they are in abstract form and not adequate to be used for TTS conversion systems. Sagisaka *et al.* formulated these rules in a good shape [3, 4], which were widely adopted in Japanese TTS conversion researches [5]. In our previous study [1], a rule-based module was developed by extending Sagisaka's rules partly.

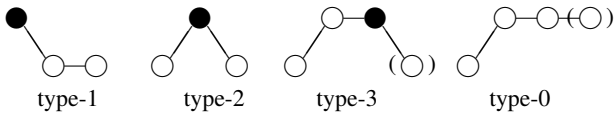

Figure 1: Accent types observable in 3-mora words of Tokyo dialect of Japanese

However, it is true that covering all the accent sandhi phenomena by rules is very difficult. In [3, 4], only the locations of primary accent nuclei were considered with the problem of secondary accent nuclei unsolved. Further, the sentences including function word concatenation were not adequately treated, either. To solve these problems, a corpus-based approach has been taken recently. In [6], n-gram models were used to develop a morphological analyzer which can produce the H/L attribute for each mora[1] of an input sentence. To take a corpus-based approach, a large corpus with accurate accent labeling is naturally required but we don't have any publicly available accent corpus. In this paper, at first, we developed a text corpus with accurate accent labeling, which will be publicly available in the near future. Using the corpus built so far, we developed a corpus-based module of predicting the accent change for adequate prosody generation. Further, the module was tuned by integrating the rule-based knowledge acquired in the previous study.

2. Word accent sandhi rules of Japanese

2.1. Word accent of Japanese

Word accent is one of the lexical attributes specific to each word and it is represented by a sequence of binary F_0 levels (H/L) in mora unit. Although it implies 2^N different accent types for N-mora words, the number of accent types for N-mora words of Tokyo dialect is reduced to N+1 due to the following properties.

1. A rapid rising or falling of F_0 has to occur between the first mora and the second one.
2. The number of the rapid falling pattern(s) of F_0 between two consecutive morae in a word is one at most.

Accent type showing a rapid downfall of F_0 immediately after the n-th mora is called type-n word accent and the n-th mora in this case is called accent nucleus. Fig. 1 shows the four accent types of 3-mora words of Tokyo dialect and their accent nuclei indicated by filled black circles. It should be noted that type-0 accent means that there is no accent nucleus and that type-0 accent and type-n accent of n-mora words are identical if they

[1]Mora is the minimum linguistic unit for speech production in Japanese, the size of which is rather similar to that of syllable.

are uttered isolatedly. The difference between the two is observed only when they are produced in connected speech. When a function word follows a type-n word, a falling pattern of F_0 is found immediately after the word. On the other hand, there is no falling patterns for type-0 words. In Fig. 1, a parenthesized circle represents the first mora of the following function word.

2.2. Word accent sandhi rules of Japanese

When a word is concatenated with another to form an accentual phrase, the resulting position of the accent nucleus of the phrase is often different from any positions of the original nuclei of the constituent words. The word accent sandhi can be categorized into three types;

1. **Shift** of the accent nucleus
   アカ + エンピツ → アカエンピツ
   red pencil
2. **Generation** of the accent nucleus
   ケイタイ + デンワ → ケイタイデンワ
   portable telephone
3. **Deletion** of the accent nucleus
   ケイザイ + テキ → ケイザイテキ
   economy (suffix) economical

The word accent sandhi in Japanese was well formulated for TTS research in [3, 4]. The following sections briefly describe the rules, which are composed of three sets of rules and several control rules over them. For each word, (a part of) three accentual attributes of concatenation manner (CM), nucleus position (NP), and concatenation type (CT) have to be prepared.

2.2.1. Concatenation of a content word and a function word to form an accentual phrase

Suppose that the concatenation of a content word of N_1 morae and type-M_1 accent and a function word (an auxiliary verb or a particle) of N_2 morae and NP being $\widetilde{M_2}$ produces an accentual phrase of N_c morae and type-M_c accent. NP is an attribute indicating the accent nucleus position in the produced accentual phrase. If the resulting accent nucleus is located as the last mora of the first word in the phrase, NP is zero. If the first mora of the second word is the accent nucleus, NP is one. It should be noted that NP can take a negative value.

If every word which can appear as the second word has its own value of NP, CM is not needed. This is because, as told above, the location of the accent nucleus is determined only by NP. In some cases, however, the accent nucleus of the first word remains after the concatenation. In these cases, the nucleus position of the phrase cannot be predicted only by the accentual attributes of the second function word. To sum up, it can be said that the accent nucleus position of an accentual phrase composed by a content word and a function word is determined by the length and the accent type of the first word and CM and NP of the second word. Table 1-(a) shows these word accent sandhi rules. As shown in the table, all of the factors above are not always required to determine the nucleus location in the phrase.

2.2.2. Concatenation of two content words

Word accent sandhi observed when concatenating two content words can be characterized by adequately setting the CM and NP values of the second *content* word. It means that these values have to be prepared for every content word. But when the second word is a verb or an adjective, the accent nucleus of the resulting phrase is always found as the last mora but one in the phrase (M_c=N_1+N_2−1). This property of Japanese requires that the values of CM and NP should be prepared only for the nouns which can occur as the second word. In this case, unlike function words described in the previous section, the CM value of the second noun word is always F4 or F5. Then, the NP value has only to be prepared for the noun word. Tab. 1-(b) shows the word accent sandhi rules in concatenating a content word and a noun. Although concatenation types (CT) are newly defined in the table, they are functionally the same as NP. C1 to C4 correspond to the NP values of M_2, 1, 0, $-N_1$ respectively. As the NP values of nouns of three morae or longer can be automatically calculated by their length and accent types, only the nouns of two morae or shorter should be considered.

Table 1: Word accent sandhi rules of Japanese
word of N_1 morae and type-M_1 accent +
word of N_2 morae and nucleus position (NP) being $\widetilde{M_2}$
→ accentual phrase of N_c morae and type-M_c accent

(a) Concatenation of a content word and a function word

concatenation manner	M_c	
	M_1=0	$M_1\neq 0$
(F1) 従属型*	M_1	
(F2) 不完全支配型*	$N_1+\widetilde{M_2}$	M_1
(F3) 融合型*	M_1	$N_1+\widetilde{M_2}$
(F4) 支配型*	$N_1+\widetilde{M_2}$	
(F5) 平板化型*	0	

(b) Concatenation of a content word and a noun

concatenation type	conditions of the 2nd word	M_c
(C1) 保存型*	$N_2 \geq 2, M_2 \neq 0, N_2$†	N_1+M_2
(C2) 生起型*	$N_2 \geq 2, M_2 = 0, N_2$†	N_1+1
(C3) 標準型*	$N_2 \leq 2$	N_1
(C4) 平板型*	$N_2 \leq 2$	0

(c) Concatenation of a prefix and a content word

concatenation type	M_c	
	M_2=0, N_2†	$M_2\neq 0$, N_2†
(P1) 一体化型*	0	N_1+M_2
(P2) 自立語結合型*	N_1+1	N_1+M_2
(P3) 分離型*	M_1	M_1 (and N_1+M_2)
(P4) 混合型*	N_1+1 (or)M_1	M_1 (and/or) N_1+M_2

∗ : In Sagisaka's original paper in Japanese, as shown here, each value of CM and CT has a meaningful name, not a label. Due to limited space, however, these values are referred to by the labels of Fx, Cx, and Px in this paper.
† : If the final syllable of the second word is comprised of two morae, N_2 should be decremented by one.

2.2.3. Concatenation of a prefix and a content word

To make an accentual phrase by attaching a suffix to a content word, the rules in Section 2.2.2 can be basically applied as they are. For a phrase composed by a prefix and a content word, new rules should be prepared, which are shown in Table 1-(c). It should be noted that, for P3 and P4, semantic analysis is sometimes required to adequately locate the accent nucleus.

In addition to the above rules, several control rules have to referred to when the above rules are used in a TTS system. Due to the limit of space, the control rules are not shown here.

3. Assignment of accent labels to a text corpus by a single labeler

In our previous study [1], the accentual attributes required by the accent sandhi rules were estimated experimentally and they were used in some TTS system developments [5]. However, covering all the accent sandhi phenomena by rules is very difficult. In the rules, only the locations of primary accent nuclei were considered with the problem of secondary accent nuclei unsolved. Further, the sentences including function word concatenation were not adequately treated, either. To solve these problems, a corpus-based approach has been adopted recently. This new approach, however, naturally requires a text corpus with accurate accent labeling but it does not exist publicly.

In the current section, the development of a text corpus with accent labeling is described in detail and its actual use in the TTS system will be shown in the following section.

3.1. What kind of labeling should be done?

If one tries to build a rule-based module to predict the accent change, for each word, he has to prepare the values of the accentual attributes described in the previous section. In the corpus-based prediction of the accent change, the values of these rather-complicated accentual attributes are not required explicitly. For example in [6], n-gram models were used to develop a morphological analyzer which can predict the H/L attribute for each mora of an input sentence. In this work, the accentual attributes of the previous section were not used at all. For training the n-gram models, only the lexical attributes, often used in the text analyzer, were referred to in addition to the H/L values of each mora of the training sentences. It should be noted that the H/L values of each mora of the constituent words when they are uttered isolatedly were not used in [6]. Even with this strategy, the prediction performance was shown to be very high. In the current paper, another statistical and machine-learning method was adopted, which is Conditional Random Fields (CRFs) [7].

CRFs are a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. Also in the case of using CRFs, the values of the accentual attributes discussed in the previous section are not needed. In the current study, the following three labels were added manually to some existing text corpora.

1. Location of the accentual phrase boundary
   A sentence utterance can be divided into several segments according to the global F_0 movement. At the beginning of each segment, F_0 rises and then, it gradually falls without a F_0 rise in the segment. The mora with the F_0 rise is the first mora of an accentual phrase and all the phrase boundaries were manually annotated. As the boundary location depends on speaking rate, the annotation was done so that a labeler could assign the boundary by looking at the reading rate indicator (See Fig. 2). The labeler was asked to read a given sentence silently according to the indicator before the assignment.

2. Location of the accent nucleus in every accentual phrase
   In an accentual phrase, according to the lexical attribute of the constituent words, one or sometimes plural rapid F_0 downfalls are observed. The mora immediately before the downfall is called accent nucleus. If plural F_0 downfalls are found in a phrase, it is considered that the first one is primary and the others are secondary accents.

3. Location of the accent nucleus in every content word when uttered isolatedly
   In this work, unlike [6], to predict the accent change, the nucleus location of each content word when uttered isolatedly was considered. The labeler was asked to indicate the nucleus position of every content word.

3.2. Selection of the single labeler

Two speakers even of the same dialect sometimes claim different accent nucleus positions for the same sentence. As phonological knowledge such as accent sandhi rules is implicit and, exactly speaking, is considered to be speaker-dependent, we decided to ask a single labeler to assign the above three labels to the whole text corpus by reading each sentence silently. As told in Sect. 2, the word accent in Japanese is mainly controlled by F_0. Then, at first, we selected 6 university students who had a good ear for the height of tone. They were members of chorus clubs and born and brought up in Tokyo. After teaching them Japanese phonology and the accent sandhi rules, we examined how sensitive they could be to linguistic sounds. In other words, we examined how well they could explicitly describe what they had in their brains implicitly. Finally, we selected a single student as labeler and asked her to assign the three kinds of labels.

As will be told in the following section, the total number of the sentences which the labeler had to deal with was more than 15 thousands. Due to the large size of the task, annotation errors may be unavoidable. Then, out of the remaining five students, we selected a few examiners, who were asked to check all the annotations. If they found some strange labels, these were fed back to the labeler, who evaluated these labels again.

3.3. Selection of the text corpus

The sentences used in the Japanese Newspaper Article Sentence database (JNAS) [8] were adopted as the text corpus. The sentences can be divided into two parts, 16,178 sentences from The Mainichi Newspapers and 503 from ATR phoneme-balanced sentences. The reasons for selecting JNAS were that all the sentences had been assigned their phonographic representation[2] and that a speech corpus for all the sentences already existed. Since the speech corpus is composed of 306 speakers, each reading a part of the corpus, it is not adequate to ask the labeler to determine the accent nucleus positions by hearing them. Further, she claimed that it was easier by reading than hearing.

3.4. Morphological analysis done on the text corpus

Every kind of content word in JNAS was separately assigned its accent nucleus position. Further, in developing a module to predict the accent sandhi using CRFs, many lexical and phonological attributes of every word of the JNAS sentences are needed. Then, morphological analysis was done on the whole sentences. Chasen [9] and UniDic [10] was adopted as morphological analyzer and dictionary. As for part-of-speech (POS), UniDic-based POS was used. The combination of Chasen and UniDic can automatically generate the phonographic representation of

[2] Japanese has two types of writing systems, phonographic (Kana) and ideographic (Kanji) systems. The sentences in newspapers are usually represented using the both systems and it is sometimes difficult to automatically determine how to convert the ideographic part into its phonographic representation.

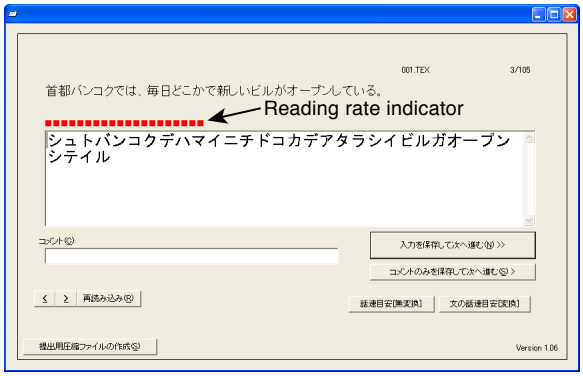

Figure 2: GUI for labeling the JNAS corpus

an input text and they showed how to read the individual sentences in JNAS. A small part of the outputs were different from the phonographic representations prepared in Sect. 3.3. For uniformity, these mismatches were manually fixed. When assigning the labels, in the case that the labeler pointed out some strange phonographic symbols of a given sentence, we gave the highest priority to the judgment of the labeler and adopted it.

3.5. Procedure of the actual accent labeling

As described in Sect. 3.1, the labeler was asked to read a given sentence silently according to the reading rate indicator (See Fig. 2). The indicator shows the rate of 7 [morae/sec] because this value is widely accepted in developing TTS systems. After reading, the labeler determined the locations of the phrase boundaries and those of the accent nuclei. As for assigning the accent nucleus position separately for each content word, a dummy word was added if necessary to follow the focused word. As told in Sect. 2.1, type-0 and type-n words are not discriminable if they are presented isolatedly. To avoid this confusion, we asked the labeler to add particle "ガ" after the given word when it was a noun and to add noun word "コト" when the given word was an adjective. The followings are examples.

首都バンコクでは，毎日どこかで新しいビルがオープンしている。
シュトバンコクデハマイニチドコカデアタラシイビルガオープンシテイル

becomes

シュ’ト／バ’ンコクデ’ハ／マ’イニチ／ド’コカデ／アタラシ’イ／ビ’ルガ／オ’ープン／シテイル.

As for the separate labeling,

シュト (ガ)
アタラシイ (コト)

becomes

シュ’ト (ガ)
アタラシ’イ (コト).

"／" means the position of the accent phrase boundary and "’" indicates that of the accent nucleus.

3.6. Discussions

As of the end of March 2007, the accent labeling of 4,166 sentences, about a fourth of the corpus, were completed and the rest of the sentences will be dealt with later. Tab. 2 shows the number of morphemes in an accentual phrase. It is found that the phrases whose word-based length is less than 5 occupy more than 90% of all the phrases. Tab. 3 shows the number of POS

Table 2: The number of morphemes in an accentual phrase

#morphemes	#occurrences	
1	5,079	(17.4%)
2	9,829	(33.6%)
3	7,902	(27.0%)
4	3,972	(13.6%)
5	1,586	(5.4%)
6	554	(1.9%)
>6	303	(1.0%)

Table 3: The number of POS patterns in the accentual phrases

POS pattern	POS pattern	#occurrences	
[名][助]	[N][P]	5,273	(18.0%)
[名]	[N]	2,639	(9.0%)
[名][名][助]	[N][N][P]	2,180	(7.5%)
[名][接尾][助]	[N][S][P]	1,409	(4.8%)
[動][助動]	[V][AV]	792	(2.7%)
[動]	[V]	788	(2.7%)
[名][名]	[N][N]	758	(2.6%)
[名][接尾]	[N][S]	739	(2.5%)
[動][助]	[V][P]	571	(2.0%)
[名][助][助]	[N][P][P]	541	(1.9%)
others		13,535	(46.3%)

名:Noun, 助:Particle, 接尾:Suffix,
動:Verb, 助動:Auxiliary Verb

patterns in all the phrases, where the top 10 frequent patterns are listed. From this table, we can say that the phrases below the top 10 occupy about a half of the phrases. In the following sections, using the corpus built so far, CRF-based statistical learning is investigated to predict the word accent sandhi.

4. CRF-based statistical learning of the word accent sandhi

4.1. Conditional Random Fields (CRFs)

CRFs are a probabilistic framework and it defines a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. In CRFs, conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$, where $\boldsymbol{y}$ and $\boldsymbol{x}$ are random variables for label and observation, is trained in the following way. Here, independent features fs are prepared about the temporal transition from y_t to y_{t+1}, called transition feature, and the generative relation between y_t and x_t, called observation feature. Let θ_f be the degree of importance of feature f and $\phi_f(\boldsymbol{x}, \boldsymbol{y})$ be the frequency of feature f being observed in the training data. Using these parameters, $P(\boldsymbol{y}|\boldsymbol{x})$ is modeled as

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y} \in Y} \left\{ \exp \sum_f \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y}) \right\}}.$$

In the training, θ_f is optimized to maximize $P(\boldsymbol{y}|\boldsymbol{x})$ for the training data. In this paper, CRF++ toolkit [11] was utilized.

4.2. What to learn with what?

In the text corpus with accent labeling, the positions of the accentual phrase boundaries and those of the accent nuclei are annotated. All the sentences are divided into accentual phrases and, on each phrase, the accent type of the constituent words is learned as $\boldsymbol{y}$ using their various lexical and phonological attributes as $\boldsymbol{x}$. It should be noted that there is a big difference

between our previous study and the current study in interpreting the resulting accentual property of each phrase generated by concatenating some words. For example, オンセーゴ┐ーセー (speech synthesis) is generated by concatenating two separate words of オ┐ンセー (speech) and ゴーセー (synthesis).

オ┐ンセー (1) + ゴーセー(0) → オンセーゴ┐ーセー (5)

In the previous study, this accent sandhi was interpreted as follows. A type-1 word and a type-0 word are concatenated to form a long compound word of type-5. In the current study, however, it is interpreted as follows.

オ┐ンセー (1) + ゴーセー(0) → オンセー(0) + ゴ┐ーセー (1)

Through the concatenation, a type-1 word and a type-0 word are transformed into type-0 and type-1. Considering the function of the word accent, this interpretation seems weird linguistically because the accent sandhi is considered to function as grouping plural words into one entity. We can say that the CRF-based statistical learning of the accent sandhi only captures the superficial transformation of the accent type of the individual words through the concatenation. The validity of this kind of approach should be carefully investigated but this paper aims to report only the performance of CRFs to predict the accent sandhi.

3,581 sentences (25,692 accentual phrases) were used to train the CRF models and the remaining 527 sentences (3,533 accentual phrases) were used to test the models.

4.3. CRF-based learning as step 1

As the first step, CRFs were examined without using the accent type of the constituent words when they are uttered isolatedly. This condition is the same as that in [6]. In the rest of the paper, the accent type of the word when uttered isolatedly is called *isolated* accent type and the accent type observed when the word is embedded in a phrase is referred to as *embedded* accent type. As observation feature, each of the followings was considered as $\boldsymbol{x}$; POS, inflection types, and the mora-based length of $w_{t-2}, w_{t-1}, w_t, w_{t+1}$ and w_{t+2}. The POS and the inflection types of UniDic are defined using multiple granularity. Following this definition, various kinds of POS and inflection types were investigated. As for transition feature, the embedded accent type of any of two consecutive words of w_{t-1} and w_t was considered. The performance of CRFs is calculated in three different ways, shown in Tab. 4. The prediction performance for all the accentual phrases, that for the phrases comprised of only two words as (noun|verb|adjective)+(auxiliary verb|particle), called as *simple* phrases henceforth, and that for the phrases including a compound noun word comprised of several consecutive nouns, called as *compound* phrases. The morpheme-based performance is also calculated for reference. Tab. 4 shows that the overall performance is 82.1%. Although it improves up to 85.9% for the simple phrases, it reduces down to 77.9% for the compound phrases[3]. It is more difficult to predict the nucleus position correctly in the compound phrases.

4.4. CRF-based learning as step 2

In addition to the observation features used above, the isolated accent type was also used here. The three kinds of performance are shown in Tab. 4. Although the overall performance and the simple performance are much improved, with the compound phrases, a slight reduction is observed. In the simple phrases, in many cases, the accent nucleus position is unchanged through concatenating the two words. On the other hand, in the compound phrases, the nucleus position is often changed through concatenating two nouns to form a compound noun. Examples are shown in Sect. 2.2 (アカエ┐ンピツ and ケイタイデ┐ンワ).

[3]For the phrases where plural nucleus positions are annotated, only the first nucleus is considered because it is the primary accent nucleus.

4.5. CRF-based learning as step 3

In the above experiments, the embedded accent nucleus position was directly predicted. Therefore, the following two cases were separately handled and modeled. A case that both the isolated nucleus and the embedded nucleus are located at the first mora and another case that they are located at the second mora. These two cases can be commonly and simply treated as "not changed" if the *relative* change of the nucleus position from isolated to embedded is predicted. In the current section, the target of the prediction was set to the relative change in the nucleus position and the following labels were prepared.

When both the isolated accent and the embedded accent had the nucleus, the labels of [0],[+1],[+2],...,[-1],[-2],... were prepared to represent the relative change of the nucleus position. When the embedded accent did not have the nucleus, the label of [none] was prepared and CRFs were trained to predict that label. When the isolated accent did not have the nucleus, the nucleus position was directly predicted as in the last section.

The performance of the relative change prediction is also shown in Tab. 4. In all the cases of all, simple, and compound, the performance is successfully increased. Especially, the increase is larger in the compound phrases. By introducing the relative change prediction, it seems that what was difficult to predict in the previous section can be adequately handled.

4.6. CRF-based learning as step 4

In this section, the training of CRFs is tuned to the accent sandhi rules described in Sect. 2. In the experiments so far, as observation feature, the generative relation between label y and lexical or phonological attribute x of observed word w was used. Referring to the accent sandhi rules, however, some kinds of the relation should be additionally considered such as that between y and x of some plural words, w_t and w_{t+1}, for example. As described in the Sect. 4.3, various kinds of the syntactic categories with multiple granularity were provided by UniDic. By carefully observing the accent sandhi rules, we prepared some word combinations to fit the CRF training to the rules. The followings are examples. [POS of w_t/POS of w_{t-1}], [POS of w_t/POS of w_{t+1}], and [fundamental lexical attributes of w_t/POS of w_{t+1}]. The fundamental lexical attributes are a set of the attributes selected adequately from the whole set of syntactic categories provided by UniDic. They included POS, inflection types, phonographic and logographic representations, and so forth.

The performance is shown in Tab. 4, again. Although a very slight reduction is found in the simple phrases, the overall performance is improved. Especially, as in the previous section, the increase is larger in the compound phrases. As described in Sect. 4.3, the nucleus prediction is more difficult in the phrases including compound nouns. We consider that the complicated phonological phenomena could be modeled by CRFs better by means of additionally introducing a set of rather complicated word combinations as better observation features.

4.7. CRF-based learning as step 5

Some additional tuning to the accent sandhi rules was investigated. In Sect. 4.5, the labels were prepared to indicate the

Table 4: The performance of the CRF-based statistical learning of the Japanese word accent sandhi

	morpheme-based	phrase-based		
		all	simple	compound
Step 1	9080 /9908 (91.6%)	2833 /3533 (82.1%)	703 /822 (85.9%)	530 /688 (77.0%)
Step 2	9272 /9908 (93.6%)	3081 /3533 (87.2%)	775 /822 (94.3%)	523 /688 (76.0%)
Step 3	9319 /9908 (94.1%)	3137 /3533 (88.8%)	791 /822 (96.2%)	553 /688 (80.4%)
Step 4	9424 /9908 (95.1%)	3214 /3533 (91.0%)	790 /822 (96.1%)	578 /688 (84.0%)
Step 5	9458 /9908 (95.5%)	3238 /3533 (91.7%)	792 /822 (96.4%)	589 /688 (85.6%)

Table 5: The performance of the CRF-based statistical learning based on the labeler's judgment

	phrase-based		
	all	simple	compound
Step 5	3307 /3533 (93.6%)	808 /822 (98.3%)	605 /688 (87.9%)

relative change of the nucleus position. In this section, the categories of these relative labels were modified to fit the module to the rules much better. As the detailed description of the modification may be tedious to readers, we show only some examples.

When the isolated accent had the nucleus, the following labels were prepared. 1) When the embedded accent did not have the nucleus, the label was [none], 2) When the embedded accent was the same as the isolated accent, the label was [same], 3) When the embedded accent nucleus was located at the last mora of the word, the label was [morae], 4) When the embedded accent type was smaller than the isolated type by 1, the label was [same-1], 5) When the embedded accent type was 1, the label was [one], 6) When the embedded accent nucleus was located at the last mora but one in the word, the label was [morae-1], 7) When the embedded accent did not correspond to any case from 1) to 6), the labels of the relative change such as [0], [+2], and [-1] were used. To assign a label to w_t, it was possible that the word could satisfy plural conditions and, in this case, the condition of the smallest number was applied. In other words, the above conditions were examined in the incremental order because the order reflected the frequency of the labels. In the last condition, the relative change labels were assigned. Before that, the labels introduced newly in this section were given to w_t. Some good readers may wonder why these cases should be treated as special cases. All of these cases were directly treated by the accent sandhi rules and, in this section, only the exceptional cases were treated by the relative change labels.

Some other tuning was also done with respect to observation features including the phonographic representation of the second mora of w_t, that of the mora located at the isolated accent nucleus position, and so forth. These features were required to fully implement the accent sandhi rules in Sect. 2 as observation features in the CRF training.

The performance is shown in Tab. 4 and only the small improvements are observed in all the three cases.

5. Discussions and conclusions

In the previous section, only the nucleus positions which the single labeler had provided were treated as correct. It is true, however, that the labeler did not reject all the other positions. We asked the labeler to judge the degree of acceptance of the accent nucleus positions predicted *incorrectly* by the CRF module. The judgment was done by using a 4-degree scale. If the nucleus positions of acceptance level 3 or 4 are re-considered as correct, the final performance is shown in Tab. 5. 93.6% of all the phrases showed the correct position and, in the simple phrases, the performance reached 98.3%. Considering that the performance of the rule-based module for the same testing data is 76.8% and 94.5% respectively for the two cases, we can claim strongly that the proposed method showed the remarkably better performance and it is very effective practically. As discussed in Sect. 4.2, however, the CRF-based implementation of the accent nucleus prediction is somewhat weird linguistically. At least, the proposed module can predict the accent change but it does not know the linguistic function of the change at all. We may have to reconsider how to train CRFs for this task.

6. References

[1] N. Minematsu, R. Kita, and K. Hirose, "Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion," Trans. IEICE, vol.E86-D, no.3, pp.550–557 (2003)

[2] "Word accent dictionary of Japanese pronunciation," published by NHK (Nippon Hoso Kyokai) (1998, in Japanese).

[3] Y. Sagisaka, and H. Sato, "Accentuation rules for Japanese text-to-speech conversion," Review of the Electrical Communication Laboratories, vol.32, no.2, pp.188-199 (1984).

[4] Y. Sagisaka, and H. Sato, "Accentuation rules for Japanese word concatenation," Trans. IECE Jpn., vol.66D, no.7, pp.849–856 (1983, in Japanese).

[5] S. Kawamoto, *et al.*, "Galatea: open-source software for developing anthropomorphic spoken dialogue agents," in *Life-like Characters, Tools, Affective Functions, and Applications*, Helmut Prendinger *et al.* (Eds.), Springer, pp.187–212 (2003)

[6] T. Nagano, S. Mori, and M. Nishimura, "An N-gram-based approach to phoneme and accent estimation for TTS," Trans. IPS Japan, vol.47, no.6, pp.1793–1801 (2006)

[7] J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. Int. Conf. Machine Learning, pp.282–289 (2001)

[8] JNAS: Japanese Newspaper Article Sentences, http://www.mibel.cs.tsukuba.ac.jp/jnas

[9] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, M. Asahara, "Japanese Morphological Analysis System: ChaSen," http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.9.pdf http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.3.3-j.pdf

[10] Y. Den, A. Yamada, H. Ogura, H. Koiso, T. Ogiso, "Japanese Morphological Analysis Dictionary: UniDic," http://download.unidic.org

[11] T. Kudo, "CRF++: Yet Another CRF Toolkit," http://crfpp.sourceforge.net